Spatiotemporal prediction of fine particulate matter using high resolution satellite images in the southeastern U.S 2003-2011

Mihye Lee[1], Itai Kloog[2], Alexandra Chudnovsky[3], Alexei Lyapustin[4], Yujie Wang[5], Steven Melly[6], Brent Coull[7], Petros Koutrakis[1], Joel Schwartz[1]

[1]Exposure, Epidemiology, and Risk Program, Department of Environmental Health, Harvard School of Public Health, Boston, MA, USA

[2]Department of Geography and Environmental Development, Ben-Gurion University of the Negev, Beer Sheva, Israel

[3]Department of Geography and Human Environment, Tel-Aviv University, Israel

[4]GEST/UMBC, NASA Goddard Space Flight Center, Baltimore, MD, USA

[5] University of Maryland Baltimore County, Baltimore, MD, USA

[6]Department of Epidemiology and Biostatistics, Drexel University School of Public Health, Philadelphia, PA, USA

[7]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

Abstract

Numerous studies have demonstrated that fine particulate matter ($PM_{2.5}$, particles smaller than 2.5 μm in aerodynamic diameter) is associated with adverse health outcomes. The use of ground monitoring stations of $PM_{2.5}$ to assess personal exposure; however, induces measurement error. Land use regression provides spatially resolved predictions but land use terms do not vary temporally. Meanwhile, the advent of satellite-retrieved aerosol optical depth (AOD) products have made possible to predict the spatial and temporal patterns of $PM_{2.5}$ exposures.

In this paper, we used AOD data with other $PM_{2.5}$ variables such as meteorological variables, land use regression, and spatial smoothing to predict daily concentrations of $PM_{2.5}$ at a 1 km^2 resolution of the southeastern United States including the seven states of Georgia, North Carolina, South Carolina, Alabama, Tennessee, Mississippi, and Florida for the years from 2003 through 2011. We divided the study area into 3 regions and applied separate mixed-effect models to calibrate AOD using ground $PM_{2.5}$ measurements and other spatiotemporal predictors.

Using 10-fold cross-validation, we obtained out of sample R^2 values of 0.77, 0.81, and 0.70 with the square root of the mean squared prediction errors (RMSPE) of 2.89, 2.51, and 2.82 $\mu g/m^3$ for regions 1, 2, and 3, respectively. The slopes of the relationships between predicted $PM_{2.5}$ and held out measurements were approximately 1 indicating no bias between the observed and modeled $PM_{2.5}$ concentrations.

Predictions can be used in epidemiological studies investigating the effects of both acute and chronic exposures to $PM_{2.5}$. Our model results will also extend the existing studies on $PM_{2.5}$ which have mostly focused on urban areas due to the paucity of monitors in rural areas.

INTRODUCTION

Since the Six Cities study[1], which showed a strong linear relationship between $PM_{2.5}$ and mortality between cities that differed by pollution level, a body of literature has reported effects of $PM_{2.5}$ on mortality and morbidity[2-4]. In many of those studies, the $PM_{2.5}$ exposures were assessed using concentration data obtained at a central monitoring site located in a jurisdiction or within a specified distance. However, this approach introduces information bias, and thus leads to attenuation of the magnitude of effects of air pollution or increases the variance of estimate[5-7]. Many studies have attempted to address this issue and to produce $PM_{2.5}$ concentrations for locations distant from the monitors[8-10]. This includes predicting $PM_{2.5}$ levels using regression models based on geographic covariates such as land use regressions or geostatistical interpolation methods such as kriging[8, 11, 12]. However, predictions from a land-use regression are limited to long-term exposures for chronic health effects studies, since the geographic covariates are mostly not time varying[13]. Moreover, if the amount of pollution due to a geographic predictor, e.g. traffic density, changes over time because of control technology, this is not easily incorporated into land use regression. Geostatistical methods also have limitations because of the low density of monitoring stations, rendering the results unreliable especially in rural areas. Meanwhile, the aerosol optical depth (AOD) values from the Moderate-Resolution Imaging Spectroradiometer (MODIS) satellite provide daily measurements for the entire earth. AOD is a measure of particles in a column of air and is related to $PM_{2.5}$[14]. With the advent of a new processing algorithm called Multi-Angle Implementation of Atmospheric Correction (MAIAC)[15], the spatial resolution of AOD has further improved from 10×10 km^2 to 1×1 km^2. Since the relationship between the AOD measurement and $PM_{2.5}$ is affected by various factors such as the optical properties of particulates, mixing height, and humidity, which vary daily, we

used a mixed-effect model with daily random slopes for daily calibration rather than a general regression. This provides better predictive performance than other studies using the satellite imagery for the $PM_{2.5}$ prediction without daily calibration[16].

In this paper, we used AOD satellite data and predictors such as meteorological variables, land use regression, and spatial smoothing to predict the daily concentration of $PM_{2.5}$ at a 1 km^2 resolution across the southeastern United States, including seven states of Georgia, North Carolina, South Carolina, Alabama, Tennessee, Mississippi, and Florida for the years 2003 through 2011.

DATA

Ground particulate matter measurements

We obtained $PM_{2.5}$ mass concentration data from Federal Reference Method (FRM) monitors operated by the U.S. Environmental Protection Agency (EPA) and monitors with a Teflon filter in the Interagency Monitoring of Protected Visual Environments (IMPROVE) program for a total of 257 monitoring sites.

Aerosol optical depth data

The MAIAC data were obtained from the National Aeronautics and Space Administration (NASA) at the resolution of 1 km^2. AOD data were delivered by tiles, which is the unit of spatial domain of MODIS image with an area of 10×10 degree at the equator. Our study used tiles h00v03, h01v02, h01v03, h01v04, h02v02, and h02v03. The data include the latitude and longitude in the WGS84 coordinate system, the corresponding AOD values, and a

quality flag. We deleted AOD values higher than 1.5 likely reflecting cloud contamination and AOD values over water bodies since the water reflects light and affects the reliability of AOD readings. The AOD value which was the closest in distance within a 1 km buffer was assigned to each $PM_{2.5}$ measurement.

To compare the new MAIAC data at a 1 km^2 resolution with the existing data at a 10 km^2 resolution, we decided to use the existing AOD data that we had retained. For the years 2000−2010, MODIS level 2 files from the Earth Observing System (EOS) Terra satellite were used to extract AOD values at a 10 km × 10 km resolution.

Meteorological data

We downloaded weather data from the National Climatic Data Center (NCDC, 2010) website. Weather variables include temperature, relative humidity, wind speed, visibility, and sea level pressure in the form of the daily mean. A total of 144 weather stations were used and we assigned the weather readings based on the closest distance on a specific data.

Normalized difference vegetation index

NASA provides normalized difference vegetation index (NDVI) data from the MODIS sensor. We aggregated NDVI measurements to a 1 km grid and a one month average. Specifically we used the Terra satellite product ID of MOD13A3.

Height of Planetary boundary layer

We obtained the daily height of planetary boundary layer (PBL) from the National Oceanic and Atmospheric Administration (NOAA) Reanalysis Data. The pixel resolution of PBL data was 32×32 km on a daily basis. To represent the daily PBL height, the 24-hr mean was used.

Land use variables

Emissions of $PM_{2.5}$, PM_{10}, and NO_x from point sources and county area level emissions, were downloaded from National Emission Inventory (NEI) data for 2005 from the website of the environmental protection agency (EPA 2005 NEI). To produce the percentage of urbanism for each satellite grid cell at 1 km^2 resolution, we used the national land cover database for 2011 (NLCD 2011) data at 30 meter resolution[17]. We reclassified land cover codes 22 (Developed, Low Intensity), 23 (Developed, Medium Intensity), and 24 (Developed, High Intensity) to 1 as an urban cell and assigned 0 for the rest of codes. The mean of binary vales was calculated for each 1 km grid cell. For the location of geographical predictors such as roads, major buildings, ports, airports, and water bodies, spatial data from ESRI Data & Maps 2004 were used (ArcGIS® and ArcMap™ by Esri, Copyright © Esri).

METHOD

Date preparation

For each day, we assigned the closest AOD readings within a 1 km buffer of PM monitors. We confined our analysis to $PM_{2.5}$ less than 80 $\mu g/m^3$ to eliminate influential outliers (25 observations among the total of 260,476 $PM_{2.5}$ measurements for 9 years). We also restricted our analysis to cells greater or equal in population to 10, since the southeastern U.S. includes less populated areas. AOD values > 0.5 which corresponded to $PM_{2.5}$ < 10 $\mu g/m^3$ were removed because it is likely they are due to cloud contamination. Data with AOD < 0.15 and $PM_{2.5}$ > 25 $\mu g/m^3$ were removed because we decided it is likely on those days that low PBL moved particles closer to ground level, deteriorating the relationship between AOD and ground-level $PM_{2.5}$ measurements.

The aim of our model lies in high-performance predication, not associational inference between the exposure and outcome such as in the epidemiological studies. Hence, our strategy was to eliminate observations with high residuals over 10 $\mu g/m^3$ as too likely to distort our predictions for most observations, and to choose a model based on maximizing cross-validated (CV) R^2. AOD values are not missing at random (for example there are more missing in the winter) which can distort the predictions. Thus, we used inverse probability weighting to account for this selection bias. Finally, the calibration between AOD and $PM_{2.5}$ can vary spatially, and daily. The daily variation is due to changes in particle size distribution, color, and vertical profile, and we address this by daily calibration and by using PBL data in the model using mixed effect models with the random intercept and slopes for day. To account for spatial differences in these daily slopes, we nested them within sub-regions, and to account for more permanent

differences between locations, we included land use terms in our model. Specifically, we fitted the following model:

$$E\left(PM_{2.5_{ij}}\right) = \left(\beta_0 + b_{0j} + b_{0jk}\right) + \left(\beta_1 + b_{1j} + b_{2jk}\right)AOD_{ij} + \left(\beta_2 + b_{2j}\right)temp_{ij} + \sum_{m=1}^{7}\beta_{1m}X_{1m_{ij}} + \sum_{n=1}^{15}\beta_{2n}X_{2n_i} + \beta_{25}AOD \times PBL \quad (1)$$

where $PM_{2.5_{ij}}$ is the $PM_{2.5}$ measurements at the monitoring site *i* on day *j*. β_0 is the intercept for the fixed effect (the population intercept) and b_{0j} is the overall random intercept which varies by day. b_{0jk} is the random intercept for day nested in each sub-region. Similarly, β_1 is the slope for the fixed effect of AOD, b_{1i} is the overall slope for the random effect of AOD for the day, and b_{2jk} is the random slope for each day nested in each sub-region. AOD is the AOD measurement that is used for the monitoring site *i* within 1 km of the site on day *j*. β_2 and b_{2j} represent the slopes for the fixed effect and the random effect of temperature, respectively. temp is the temperature that is measured by the closest weather monitor to the site *i* on day *j*. β_{1m} is the slopes for the fixed effect of spatiotemporal variables. X_{1mij} is the matrix of m^{th} spatiotemporal covariates on the site *i* and day *j* other than temperature and consists of 7 variables: dew point temperature, sea level pressure, visibility, wind speed, absolute humidity; NDVI in the corresponding month; and PBL. β_{2n} is the slopes for the fixed effect of spatial variables. X_{2ni} is the matrix of 15 spatial covariates for the i^{th} site which includes the percent

urbanicity, elevation, the density of major roads, population within 10 km diameter, $PM_{2.5}$ emissions at county level, $PM_{2.5}$ emissions from point sources, PM_{10} emission from point sources, NOx emission from point sources, canopy surface in 2001, distance to the closest A1 roads, distance to the closest airport, distance to the closest port, distance to the closest railroad, distance to a closest road, and distance to the major building. Observations with residuals over 10 $\mu g/m^3$ were re-visited and we determined their validity by comparing $PM_{2.5}$ readings from the surrounding monitors and the previous day and the next day. If we determined them to be erroneous, we assigned the readings from the closest monitoring station within 15 km.

Model

Due to the vast study area, a single model was not able to achieve the best performance in prediction. The southeastern U.S. consists of various areas with different topography, climate (tropical in Florida), and geographic features such as swamps and forests. Therefore, we decided to split the study area into three regions and to fit separate models for each region and implement nested random coefficients for sub-regions within each region (Figure 1). Region 1 consist of Tennessee, Mississippi, Alabama, and Georgia. Region 2 covers North Carolina, South Carolina, and Georgia. Lastly, region 3 covers Florida, Mississippi, Alabama, Georgia, and South Carolina.

AOD measurement cannot be made due to various factors such as cloud or snow cover. We hypothesized that the cloud formation and snow cover is affected by weather conditions including temperature, wind speed, sea level pressure, elevation and the season. Therefore, to adjust the non-random missingness of AOD, we modeled inverse probability weights (IPW) and

applied them to the first stage models. Specifically, we fitted the following logistic model for the missingness of AOD measurements.

$$\mathrm{E}\big(logit(p)\big) = \beta_0 + \beta_1 temp_{ij} + \beta_2 WS_{ij} + \beta_3 SLP_{ij} + \beta_4 elev_i + \beta_5 mon_j, \quad (2)$$

where temp is temperature of cell *i* on day *j*, WS_{ij} is wind speed of cell *i* on day *j*, SLP_{ij} is the sea level pressure of cell *i* on day *j*, elev is the elevation of cell *i*, and mon is the corresponding month that day *j* falls in.

Using the probability of the outcome (missing or not), we computed the inverse probability as, $\frac{1}{p}$. Next, we normalized IPW values by dividing them by their mean. These were applied to the subsequent models as a weight.

Each of the models corresponding to the three regions was evaluated using a 10 fold cross-validation to avoid over-fitting. We adopted a different approach in cross-validation which other similar studies performed record-based cross-validation. We conducted site-based cross-validation since we believed that cross-validation by monitoring stations was more appropriate so that it assesses the capabilities of the models to predict spatial variability. Firstly, we made a randomly ordered list of monitoring stations in each region. The station list then was split into 10 subsets. In turn, 90 % of monitoring stations were used to fit the model and 10 % of stations were used to test the model performance. This cross-validation were conducted for 10 times for each region. The site-based 10-fold cross-validated R^2 was used for finalizing the models rather than modeled R^2 as well as for assessing the model performance and for avoiding over-fitting. As a result, we ended up the following models based on the highest R^2 from the 10-fold cross-validation.

In region 1, we fitted the following model for each year with the IPW:

$$\mathrm{E}\left(PM_{2.5_{ij}}\right) = (\beta_0 + b_{0j} + b_{0jk}) + (\beta_1 + b_{1j} + b_{1jk})AOD_{ij} + \beta_2 temp_{ij} + \beta_3 dewp_{ij} + \beta_4 slp_{ij} + \beta_5 wdsp_{ij} + \beta_6 visib_{ij} + \beta_7 ah_{ij} + \beta_8 NDVI + \beta_9 elev_i + \beta_{10} pbl_{ij} + \beta_{11} urb_i + \beta_{12} emission_i + \beta_{13} PM10_i + \beta_{14} NOX_i \quad (3)$$

where $PM_{2.5_{ij}}$ is the $PM_{2.5}$ measurements at the monitoring site *i* on day *j*. β_0 denotes the fixed effect intercept term (population intercept) and b_{0j} is the random effect intercept varies randomly from one day to another. b_{0jk} is the random intercept for day nested in each sub-region. Similarly, β_1 is the slope for the fixed effect of AOD, b_{1i} is the slope for the random effect of AOD for each day, and b_{2jk} is the random slope for each day nested in each sub-region.

AOD is the AOD measurement that is used for the monitoring site *i* within 1 km of the site on day *j*. temp is the temperature that is measured by the closest weather monitor to the site *i* on day *j*. dewp is the dew point that is measured by the closest weather monitor to the site *i* on day *j*. slp is the sea level pressure in millibars that is measured by the closest weather monitor to the site *i* on day *j*. wdsp is the wind speed in knots that is measured by the closest weather monitor to the site *i* on day *j*. visib is the visibility in miles that is measured by the closest weather monitor to the site *i* on day *j*. elev is the elevation of the site *i*. pbl is the height of the planetary boundary layer at the site *i* on day *j*. urb is the percentage of urbaness at the site *i*. emission is the annual emission of $PM_{2.5}$ in ton from the closest point source such as an industrial factory. PM10 is the annual emission of PM_{10} in ton from the closest point source such as an industrial factory. NOX is the annual emission of NO_X in ton from the closest point source such as an industrial factory.

In region 2, we fitted the following model for each year with the IPW:

$$\mathrm{E}\left(PM_{2.5_{ij}}\right) = \left(\beta_0 + b_{0j} + b_{0jk}\right) + \left(\beta_1 + b_{1j} + b_{1jk}\right)AOD_{ij} + \beta_2 temp_{ij} + \beta_3 dewp_{ij} + \beta_4 slp_{ij} + \beta_5 wdsp_{ij} + \beta_6 visib_{ij} + \beta_7 ah_{ij} + \beta_8 NDVI + \beta_9 elev_i + \beta_{10} pbl_{ij} + \beta_{11} urb_i + \beta_{12} emission \quad (4)$$

For the third region, we fitted the following model for each year with the IPW:

$$\mathrm{E}\left(PM_{2.5_{ij}}\right) = \left(\beta_0 + b_{0j} + b_{0jk}\right) + \left(\beta_1 + b_{1j} + b_{1jk}\right)AOD_{ij} + \beta_2 temp_{ij} + \beta_3 dewp_{ij} + \beta_4 slp_{ij} + \beta_5 wdsp_{ij} + \beta_6 visib_{ij} + \beta_7 ah_{ij} \quad (5)$$

Besides the overall R^2 from the 10-fold cross-validation, we estimated a spatial R^2 by regressing the annual mean of observed $PM_{2.5}$ against that of predicted one for each site. To assess the precision of the predictions, root mean squared prediction error (RMSPE) was generated by taking the square root of the mean of squared prediction residuals. A temporal R^2 was calculated by regressing the difference between the actual $PM_{2.5}$ measurement on a specific day and the annual mean for each site against the equivalent for the predicted values from the model.

Once we finalized the calibration models by three regions as above, we predicted $PM_{2.5}$ levels based on the coefficients for AOD values and other temporal and spatial variables.

For the areas and days with AOD missing, we interpolated those cells using the surrounding cells that had AOD values and thus had predictions in the second stage. Specifically, we applied the following model with the IPW.

$$\left(PredPM_{2.5_{ij}}\right) = \left(\beta_0 + b_{0j} + b_{0jk}\right) + s(lat_i, long_i) + (\beta_1 + b_{1ik})MPM_{ij} + \beta_2 bimon_{ij} + \beta_3 pbl_{ij} + \beta_4 ah\_gm3_{ij} + \beta_5 elev_{ij} + \beta_6 mpm \times bimon_{ij} + \beta_7 mpm \times pblh_{ij}$$

,(6)

where $PredPM_{ij}$ is the predicted $PM_{2.5}$ level at a grid cell *i* on a day *j* in stage 2. lat_i and $long_i$ are the latitude and longitude coordinates of the cell *i*, respectively; and s() is a smooth function of thin plate splines. MPM_{ij} is the mean $PM_{2.5}$ measured at monitoring stations within a 100 km buffer for the cell *i* on day *j*.

Since the purpose of the analysis of the 10 km data is to compare the performance of two data, we conducted the first stage model only. During the modeling, we applied same procedures as above with the same model with same variables, calibration, and IPW to make a fair comparison.

As for software, MATLAB 2014b was used to extract the AOD readings from the raw satellite image in the HDF format and ArcGIS Desktop 10.2.2 was used along with python scripting for data preparation. Models were implemented by using the R 3.02 and SAS 9.3 (Statistical Analysis System).

RESULTS

A total of 257 monitoring stations were used for the study. Figure 1 shows the study area and the locations of $PM_{2.5}$ monitors. The study area with the thick boundary line covers most of the seven states except for the small area of western Mississippi due to the lack of the total spatial domain consisting of AOD tiles. The numbers from 1 to 3 in big bold font indicate the study area region. Region 1 mainly consists of the states of Tennessee, and the upper part of

Mississippi, Alabama, and Georgia, and contains 61 monitoring stations (0.0003 monitor/km^2). Region 2 includes most of North Carolina, and major parts of South Carolina, and Georgia with 88 monitors. Region 2 is most densely populated by PM monitoring stations (0.00038 monitor/km^2). Region 3 covers the most southern part, including Florida and the southern part of Mississippi, Alabama, Georgia, and South Carolina. Although region 3 has the largest number of monitors of 108, due to its vast area, the spatial distribution of PM monitoring stations is most scattered among the three regions (0.00026 monitor/km^2).

Table 1 shows the descriptive statistics for $PM_{2.5}$ measurements from monitoring stations and AOD measurements by MAIAC algorithm in the southeastern U.S. by year from 2003 to 2011. The annual average of $PM_{2.5}$ has steadily decreased from 12.2 in 2003 to 9.8 $\mu g/m^3$ in 2011. The standard deviation has also decreased from 6.5 to 5.3 $\mu g/m^3$. The mean AOD readings were on the order of 0.20 (dimensionless) over 9 years.

Figure 2 shows the spatial distribution of $PM_{2.5}$ concentrations in the study area, represented by the average $PM_{2.5}$ levels by monitors during the study period (2003-2011). Monitoring stations in big cities such as Atlanta, Nashville, Charlotte, and Birmingham recorded the highest average $PM_{2.5}$ level. Monitors at intersections of major highways also showed the high level of $PM_{2.5}$. Among the seven study states, Florida showed the lowest $PM_{2.5}$ level.

Our model showed a highly significant association between $PM_{2.5}$ and AOD after controlling for other covariates and spatiotemporal predictors. Table 2 presents results from the stage 1 model where the calibration of AOD and other spatiotemporal predictors were done by each year and region. The R^2 numbers are from the 10-fold cross-validation based on the sampling of monitors not observations regardless of monitors. The predictive power of the models differed by region. Region 2 showed the highest overall R^2 of 0.81 with the year-to-year

variation ranging from 0.78 in 2008 to 0.85 in 2007. Region 3 showed the lowest performance with an average cross-validated R^2 of 0.70 (minimum of 0.63 occurred in 2011 and maximum of 0.75 occurred in 2003 and 2005). For region 1, an average cross-validated R^2 was 0.77 and ranged from 0.65 in 2010 to 0.83 in 2005. The slopes between the observed $PM_{2.5}$ versus the modeled $PM_{2.5}$ were close to 1 for all the regions, suggesting a good agreement between the model results and actual measurements and the thus low bias. Region 2 exhibited the lowest average root mean square prediction error (RMSPE) of 2.51 $\mu g/m^3$, followed by region 3 with 2.82 $\mu g/m^3$ and region 1 with 2.87 $\mu g/m^3$. The RMSPE for the spatial component was much lower at 0.82 $\mu g/m^3$ in region 2. In general, the models performed better temporally than spatially. The temporal R^2 values were higher than the spatial ones except for region 3. For the temporal result, the mean R^2 was 0.80, 0.82, and 0.69 for regions 1, 2, and 3, respectively. For the spatial model the mean R^2 was 0.69, 0.63, and 0.76 by region order.

The output prediction model based on the third model gave very similar results (Table 3). The third column represents the R^2 for the prediction from stage 2 (prediction for the gird cells and days that AOD readings were available) and the last column illustrates those for the comparison with actual $PM_{2.5}$ observations. The final prediction showed high predictive power, from 0.89 (region 2) to 0.86 (region 3).

To graphically represent the predictions, Figure 3 displays the prediction results in the form of annual average in 2003 where reveals higher $PM_{2.5}$ levels for highways and the main cities. The spatial pattern of predictions matches well with the one of the measured $PM_{2.5}$ represented in Figure 2. There was no systematic spatial patterns of residuals during the study period (Figure 4).

Compared to the existing AOD data at a 10 × 10 km resolution, the MAIAC data at a 1 × 1 km resolution showed the better performance (Table 4). Only except for slight decrease in the mean of 10-fold cross-validated R^2 for Region 1 from 0.78 to 0.77, the MAIAC data showed the higher R^2 values. Especially, the performance in Region 3 drastically improved from 0.62 to 0.70. The new data also had lower errors than the existing one. RMSPE values have decreased from 3.27 to 2.89 $\mu g/m^3$ for Region 1, from 2.90 to 2.51 $\mu g/m^3$ for Region 2, from 3.64 to 2.82 $\mu g/m^3$ for Region 3. Other indicators such as Spatial R^2 and temporal R^2 have also improved when using the 1 km AOD data.

DISCUSSION

In this paper, we predicted $PM_{2.5}$ levels across the southeastern U.S. at a 1 km resolution using the MODIS satellite imagery derived by the newly developed algorithm MAIAC. Compared to the AOD data at a 10 km resolution, the MAIAC data at a 1 km resolution showed the better performance. Furthermore, higher resolution enabled the more precise exposure assessment for $PM_{2.5}$ at a finer scale such as the street-level address.

These results will enable epidemiological studies to evaluate the association between $PM_{2.5}$ and its health effects with reduced measurement error in exposure. We also anticipate study areas may extend to rural areas in the southeastern U.S., which were formerly restricted to urban areas due to the distance to monitoring stations. Considering that $PM_{2.5}$ measurements are not always daily, our model interpolates the temporal break using the daily satellite imagery and a smoothing technique as well as spatial predictions. This approach enables epidemiological studies to examine both acute and chronic effects.

Model performance varied by region. Region 2 mainly covering North Carolina revealed the highest performance (0.81) and region 3, covering the most southern part, such as Florida, had the lowest performance (0.70). One possible explanation is that the spatial density of monitoring stations affects model. Region 2 has the most abundant monitoring stations compared to its area, whereas region 3 lacks monitoring stations for its extensive area. This appeared to affect the results by providing fewer pairs to fit the model. Another explanation may be that region 2 is relatively more urbanized compared to region 3 with more land use factors which could be taken into account. This suggestion parallels with our experience during the analysis that the calibration model based on the highest R^2 for region 2 has more land use variables than that for region 3. Lastly, the quality of AOD from the MODIS instrument and the MAIAC algorithm should be considered. Visual analysis (data not present) by AOD swath revealed that the performance of AOD differed by tile of satellite imagery. Tile h01v02 that covers North Carolina showed the best performance, whereas tiles around Alabama (h00v03 and h01v03) showed the poorest performance. To improve model performance, other AOD products from other algorithms such as AOD data from Deep Blue algorithm[18] at 10 km resolution can be incorporated which is used for bright surfaces. More studies are needed to determine which factors play a role in the prediction of $PM_{2.5}$ using satellite imagery and to further improve the performance.

Compared to the existing studies on the similar area[19-21], our study shows higher R^2 and less errors. After predicting $PM_{2.5}$ levels at a 10 km resolution for the similar area for the year 2003[21], Hu et al[19]. examined the feasibility of the 1 km resolution MAIAC AOD data by comparing with the 10 km data. In their study, the performance of the MAIAC AOD data was comparable to the existing MODIS data but showed slightly lower performance. Our study

demonstrated the MAIAC AOD can outperform the existing 10 km data by using various approaches on the top of the advantage of the higher resolution. The study resulted in an R^2 of 0.64 and RMSPE of 3.93 $\mu g/m^3$ for the MAIAC data in stage 1. In our model, the lowest R^2 in 2003 was 0.72 with a RMSPE 3.51 $\mu g/m^3$. Recently, they expanded their study period for the same area[20] from a single year of 2003 to the multiple years from 2001 to 2010. Our study area covers vast additional areas in the southeastern U.S. by adding Florida, Mississippi, and the complete parts of other states. Adopting different approaches than their study, our study shows higher R^2 values and lower RMSPE. The total mean of10-fold cross-validated R^2 was 0.76 compared to the existing study 0.72 and that of RMSPE from our study was 2.74 compared to 3.72 $\mu g/m^3$. Considering that our study area includes the most southern area such as Florida which showed the lowest performance with a big difference and we applied site-based cross-validation rather than observation-based cross-validation which produces higher R^2, the actual improvement is expected to be bigger.

In conclusion, we have demonstrated that the use of satellite imagery and other land use variables with a mixed-effect model produces reliable predictions of daily $PM_{2.5}$ for the large area of the southeastern United States. By incorporating land use terms and spatial smoothing, our models perform much better than previous studies. Therefore, our model results can be used in various epidemiological studies investigating the effects of $PM_{2.5}$ allowing one to assess both acute and chronic exposures with the implication of a new application. Our model results will extend the existing studies on $PM_{2.5}$ mainly targeted only for urban areas tied to the lack of monitors into new areas which used not to be studied such as rural areas.

Acknowledgement: This publication was made possible by USEPA grant RD 83479801. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

346

REFERENCES

1. Dockery DW, Pope CA, Xu X, et al. An Association between Air Pollution and Mortality in Six U.S. Cities. *N Engl J Med*. 1993;329(24):1753-1759. doi: 10.1056/NEJM199312093292401.

2. Pope CA,3rd. Epidemiology of fine particulate air pollution and human health: biologic mechanisms and who's at risk?. *Environ Health Perspect*. 2000;108 Suppl 4:713-723.

3. Pope CA,3rd, Burnett RT, Thun MJ, et al. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*. 2002;287(9):1132-1141.

4. Barnett AG, Williams GM, Schwartz J, et al. The effects of air pollution on hospitalizations for cardiovascular disease in elderly people in Australian and New Zealand cities. *Environ Health Perspect*. 2006;114(7):1018-1023.

5. Rhomberg LR, Chandalia JK, Long CM, Goodman JE. Measurement error in environmental epidemiology and the shape of exposure-response curves. *Crit Rev Toxicol*. 2011;41(8):651-671. doi: 10.3109/10408444.2011.563420 [doi].

6. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med*. 1998;55(10):651-656.

7. Goldman GT, Mulholland JA, Russell AG, et al. Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. *Environ Health*. 2011;10:61-069X-10-61. doi: 10.1186/1476-069X-10-61 [doi].

8. Ryan PH, LeMasters GK. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhal Toxicol*. 2007;19 Suppl 1:127-133. doi: 782016666 [pii].

9. de Hoogh K, Wang M, Adam M, et al. Development of Land Use Regression Models for Particle Composition in Twenty Study Areas in Europe. *Environ Sci Technol*. 2013;47(11):5778-5786. doi: 10.1021/es400156t.

10. Beckerman BS, Jerrett M, Martin RV, van Donkelaar A, Ross Z, Burnett RT. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. *Atmos Environ*. 2013;77(0):172-177. doi: http://dx.doi.org/10.1016/j.atmosenv.2013.04.024.

11. Wang R, Henderson SB, Sbihi H, Allen RW, Brauer M. Temporal stability of land use regression models for traffic-related air pollution. *Atmos Environ*. 2013;64(0):312-319. doi: http://dx.doi.org/10.1016/j.atmosenv.2012.09.056.

12. Whitworth KW, Symanski E, Lai D, Coker AL. Kriged and modeled ambient air levels of benzene in an urban environment: an exposure assessment study. *Environ Health*. 2011;10:21-069X-10-21. doi: 10.1186/1476-069X-10-21 [doi].

13. Kloog I, Koutrakis P, Coull BA, Lee HJ, Schwartz J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical

depth measurements. *Atmos Environ*. 2011;45(35):6267-6275. doi: http://dx.doi.org.ezp-prod1.hul.harvard.edu/10.1016/j.atmosenv.2011.08.066.

14. Alston EJ, Sokolik IN, Kalashnikova OV. Characterization of atmospheric aerosol in the US Southeast from ground- and space-based measurements over the past decade. *Atmospheric Measurement Techniques*. 2012;5(7):1667-1682. doi: 10.5194/amt-5-1667-2012.

15. Lyapustin A, Wang Y, Laszlo I, et al. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *Journal of Geophysical Research: Atmospheres*. 2011;116(D3):- D03211. doi: 10.1029/2010JD014986.

16. Lee HJ, Liu Y, Coull BA, Schwartz J, Koutrakis P. A novel calibration approach of MODIS AOD data to predict PM$_2.5$ concentrations. *Atmospheric Chemistry and Physics*. 2011;11(15):7991-8002. doi: 10.5194/acp-11-7991-2011.

17. Jin S, Yang L, Danielson P, Homer C, Fry J, Xian G. A comprehensive change detection method for updating the National Land Cover Database to circa 2011. *Remote Sens Environ*. 2013;132(0):159-175. doi: http://dx.doi.org/10.1016/j.rse.2013.01.012.

18. Li X, Xia X, Wang S, Mao J, Liu Y. Validation of MODIS and Deep Blue aerosol optical depth retrievals in an arid/semi-arid region of northwest China. *Particuology*. 2012;10(1):132-139. doi: http://dx.doi.org/10.1016/j.partic.2011.08.002.

19. Hu X, Waller LA, Lyapustin A, et al. Estimating ground-level PM2.5 concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sens Environ*. 2014;140(0):220-232. doi: http://dx.doi.org/10.1016/j.rse.2013.08.032.

20. Hu X, Waller LA, Lyapustin A, Wang Y, Liu Y. 10-year spatial and temporal trends of PM$_2.5$ concentrations in the southeastern US estimated using high-resolution satellite data. *Atmospheric Chemistry and Physics*. 2014;14(12):6301-6314. doi: 10.5194/acp-14-6301-2014.

21. Hu X, Waller LA, Al-Hamdan MZ, et al. Estimating ground-level PM2.5 concentrations in the southeastern U.S. using geographically weighted regression. *Environ Res*. 2013;121(0):1-10. doi: http://dx.doi.org/10.1016/j.envres.2012.11.003.

TABLES

Table 1. Descriptive statistics of $PM_{2.5}$ ($\mu g/m^3$) and MAIAC AOD

Year	Mean PM (S.D.)	Mean AOD (S.D.)
2003	12.2 (6.5)	0.18 (0.18)
2004	12.6 (6.6)	0.18 (0.17)
2005	13.1 (7.3)	0.20 (0.19)
2006	12.6 (6.6)	0.20 (0.19)
2007	12.4 (7.5)	0.21 (0.21)
2008	10.8 (5.6)	0.18 (0.16)
2009	9.4 (4.6)	0.17 (0.15)
2010	10.2 (4.9)	0.17 (0.15)
2011	9.8 (5.3)	0.20 (0.18)

S.D., standard deviation

Table 2. Result of site-based 10-fold cross-validation from stage 1 model using 1 km^2 data

Year	Region	R^2 (CV)	Slope (CV)	RMSPE ($\mu g/m^3$)	Spatial R^2	Temporal R^2	Spatial RMSPE
	1	0.72	0.93	3.51	0.50	0.78	1.86
2003	2	0.83	0.98	2.67	0.59	0.84	1.03
	3	0.75	1.01	2.62	0.81	0.74	0.93
	1	0.79	0.97	2.92	0.94	0.80	1.07
2004	2	0.80	0.99	2.77	0.52	0.81	0.79
	3	0.74	0.99	2.83	0.77	0.74	0.86
	1	0.83	0.99	3.23	0.86	0.84	1.12
2005	2	0.80	0.97	3.12	0.81	0.81	0.93
	3	0.75	0.99	3.10	0.73	0.75	1.19
	1	0.80	0.98	2.99	0.53	0.83	1.26
2006	2	0.84	0.99	2.70	0.70	0.85	0.86
	3	0.74	1.00	2.69	0.67	0.75	1.15
	1	0.79	0.98	3.19	0.67	0.82	1.34
2007	2	0.85	0.99	2.54	0.59	0.86	0.84
	3	0.70	1.02	3.29	0.77	0.69	1.25
	1	0.78	0.99	2.71	0.74	0.80	0.99
2008	2	0.78	0.98	2.48	0.60	0.79	0.79
	3	0.69	1.00	2.74	0.85	0.65	0.99
	1	0.76	0.98	2.30	0.81	0.78	0.83
2009	2	0.78	0.99	2.05	0.81	0.79	0.78
	3	0.66	1.02	2.60	0.80	0.64	0.87
	1	0.65	0.95	2.80	0.33	0.71	1.33
2010	2	0.80	0.99	2.09	0.46	0.81	0.68
	3	0.66	1.00	2.51	0.69	0.66	1.11
	1	0.79	0.98	2.40	0.80	0.80	0.86
2011	2	0.78	0.98	2.21	0.55	0.79	0.69
	3	0.63	0.99	2.97	0.75	0.61	0.98
	1	0.77	0.97	2.89	0.69	0.80	1.18
Mean	2	0.81	0.99	2.51	0.63	0.82	0.82
	3	0.70	1.00	2.82	0.76	0.69	1.04

Table 3. R^2 from stage 3 model

Year	Region	R^2 Pred2	R^2 PM_{25}
	1	0.83	0.90
2003	2	0.86	0.91
	3	0.61	0.85
	1	0.83	0.88
2004	2	0.84	0.90
	3	0.64	0.85
	1	0.83	0.91
2005	2	0.84	0.90
	3	0.65	0.87
	1	0.86	0.89
2006	2	0.87	0.91
	3	0.59	0.86
	1	0.83	0.90
2007	2	0.84	0.91
	3	0.62	0.88
	1	0.83	0.87
2008	2	0.82	0.88
	3	0.65	0.90
	1	0.81	0.86
2009	2	0.80	0.86
	3	0.61	0.83
	1	0.75	0.83
2010	2	0.81	0.89
	3	0.60	0.85
	1	0.85	0.89
2011	2	0.81	0.88
	3	0.61	0.87
	1	0.82	0.88
Mean	2	0.83	0.89
	3	0.62	0.86

Table 4. Result of site-based 10-fold cross-validation from stage 1 model using 10 km^2 data

Year	Region	R^2 (CV)	Slope (CV)	RMSPE ($\mu g/m^3$)	Spatial R^2	Temporal R^2	Spatial RMSPE
2000	1	0.84	0.99	3.76	0.64	0.85	1.09
2000	2	0.80	0.98	3.40	0.54	0.82	1.42
2000	3	0.72	1.00	4.09	0.64	0.74	1.73
2001	1	0.78	0.98	3.68	0.43	0.80	1.55
2001	2	0.79	0.99	3.18	0.52	0.80	1.14
2001	3	0.65	0.98	3.74	0.57	0.69	1.59
2002	1	0.79	0.97	3.60	0.62	0.80	1.15
2002	2	0.77	0.99	3.14	0.36	0.79	1.14
2002	3	0.63	0.96	3.72	0.58	0.63	1.40
2003	1	0.77	0.99	3.30	0.25	0.79	1.29
2003	2	0.84	0.99	2.79	0.30	0.86	1.03
2003	3	0.60	0.96	3.56	0.52	0.61	1.54
2004	1	0.75	0.97	3.30	0.36	0.77	1.25
2004	2	0.78	0.99	3.15	0.47	0.79	0.93
2004	3	0.69	0.97	3.60	0.63	0.71	1.40
2005	1	0.82	0.98	3.59	0.38	0.84	1.33
2005	2	0.81	0.99	3.20	0.58	0.83	1.06
2005	3	0.68	0.99	3.90	0.62	0.71	1.69
2006	1	0.79	1.00	3.36	0.60	0.80	1.13
2006	2	0.82	0.99	3.03	0.46	0.83	1.08
2006	3	0.62	0.97	3.45	0.55	0.64	1.60
2007	1	0.77	0.97	3.71	0.64	0.78	1.22
2007	2	0.82	0.99	2.88	0.66	0.83	0.73
2007	3	0.59	0.98	4.31	0.57	0.59	1.67
2008	1	0.77	0.98	2.79	0.46	0.79	0.99
2008	2	0.77	0.98	2.60	0.61	0.78	0.86
2008	3	0.58	0.96	3.36	0.59	0.57	1.42
2009	1	0.77	0.99	2.33	0.63	0.78	0.80
2009	2	0.75	0.99	2.15	0.78	0.77	0.76
2009	3	0.49	1.00	3.25	0.64	0.48	1.34
2010	1	0.70	0.99	2.57	0.20	0.73	0.96
2010	2	0.75	1.00	2.38	0.53	0.76	0.77
2010	3	0.55	0.99	3.02	0.68	0.54	1.38
Mean	1	0.78	0.98	3.27	0.47	0.79	1.16
	2	0.79	0.99	2.90	0.53	0.81	0.99
	3	0.62	0.98	3.64	0.60	0.63	1.52

FIGURES

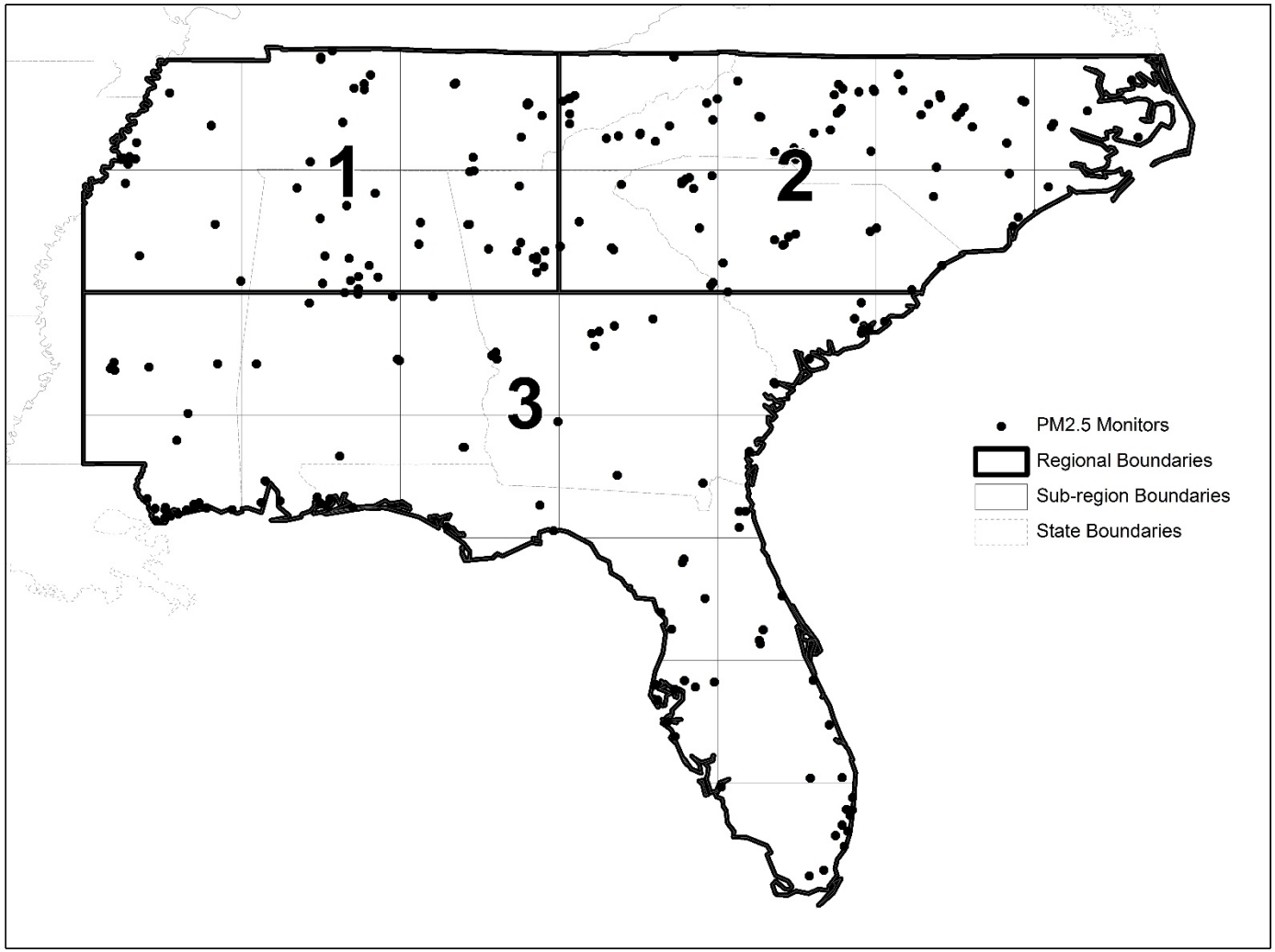

Figure 1. Study area and the locations of $PM_{2.5}$ monitoring stations

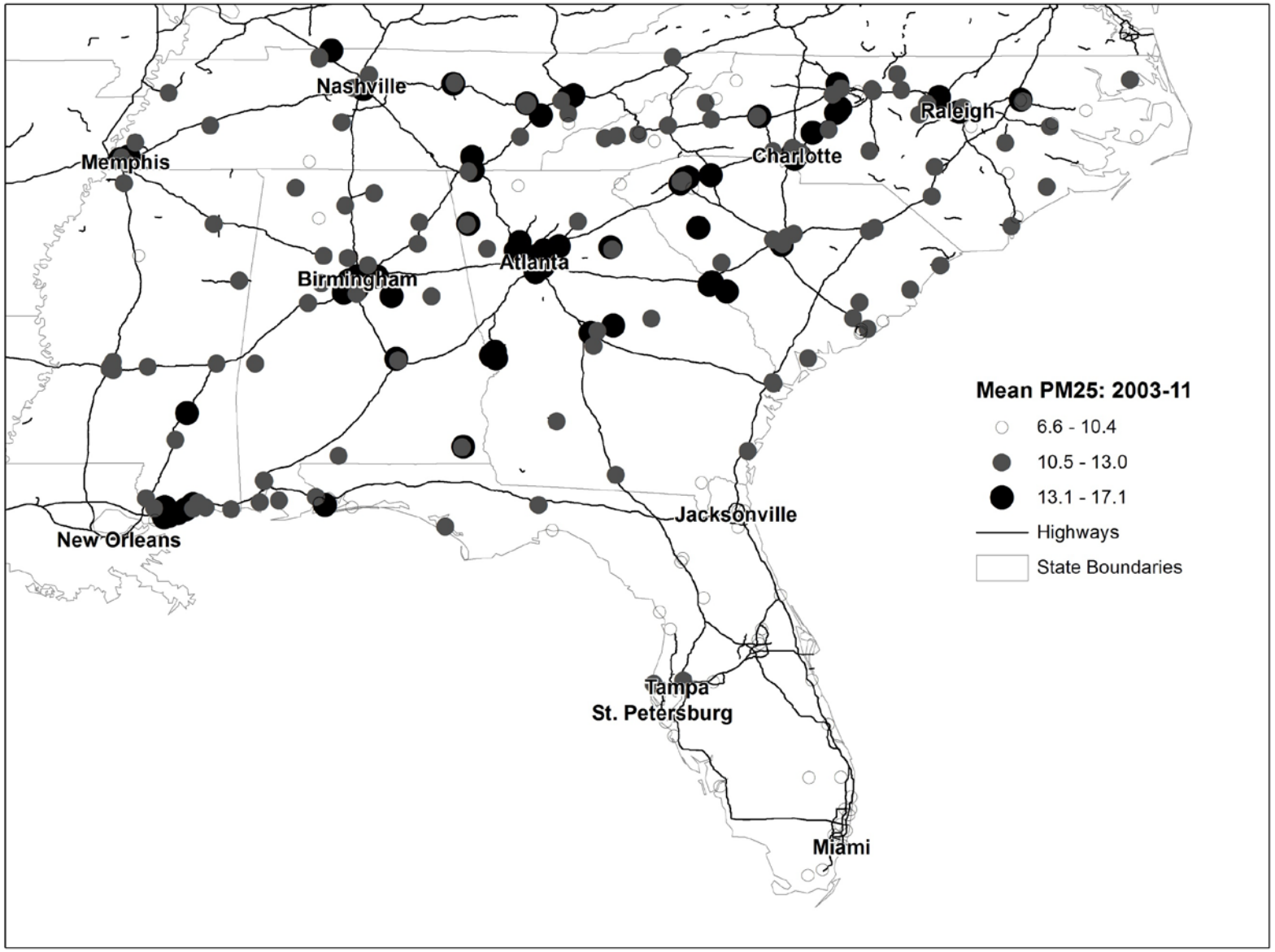

Figure 2. Spatial Distribution of $PM_{2.5}$ concentrations between 2003 and 2011

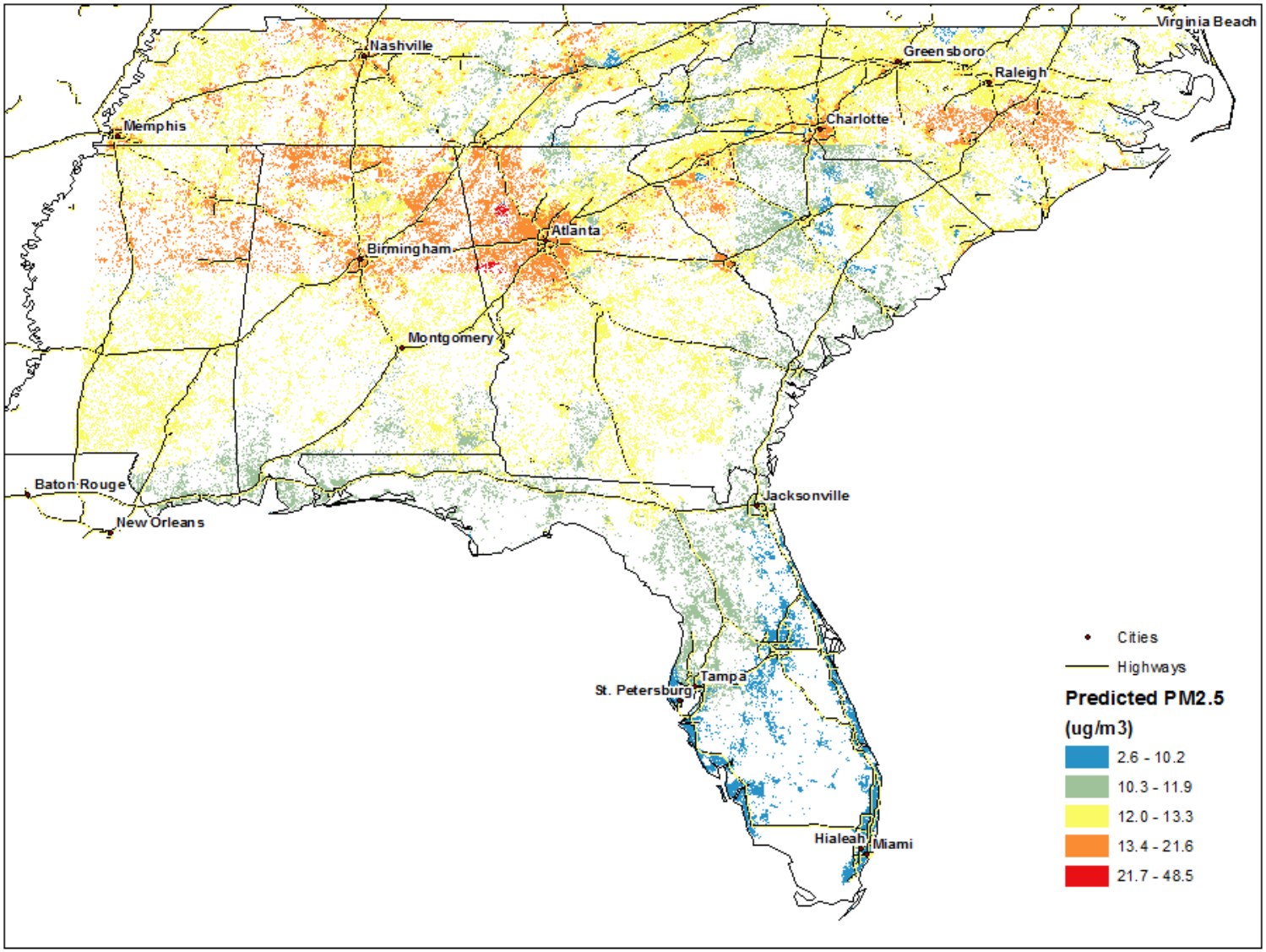

Figure 3. Predicted $PM_{2.5}$ level in 2003

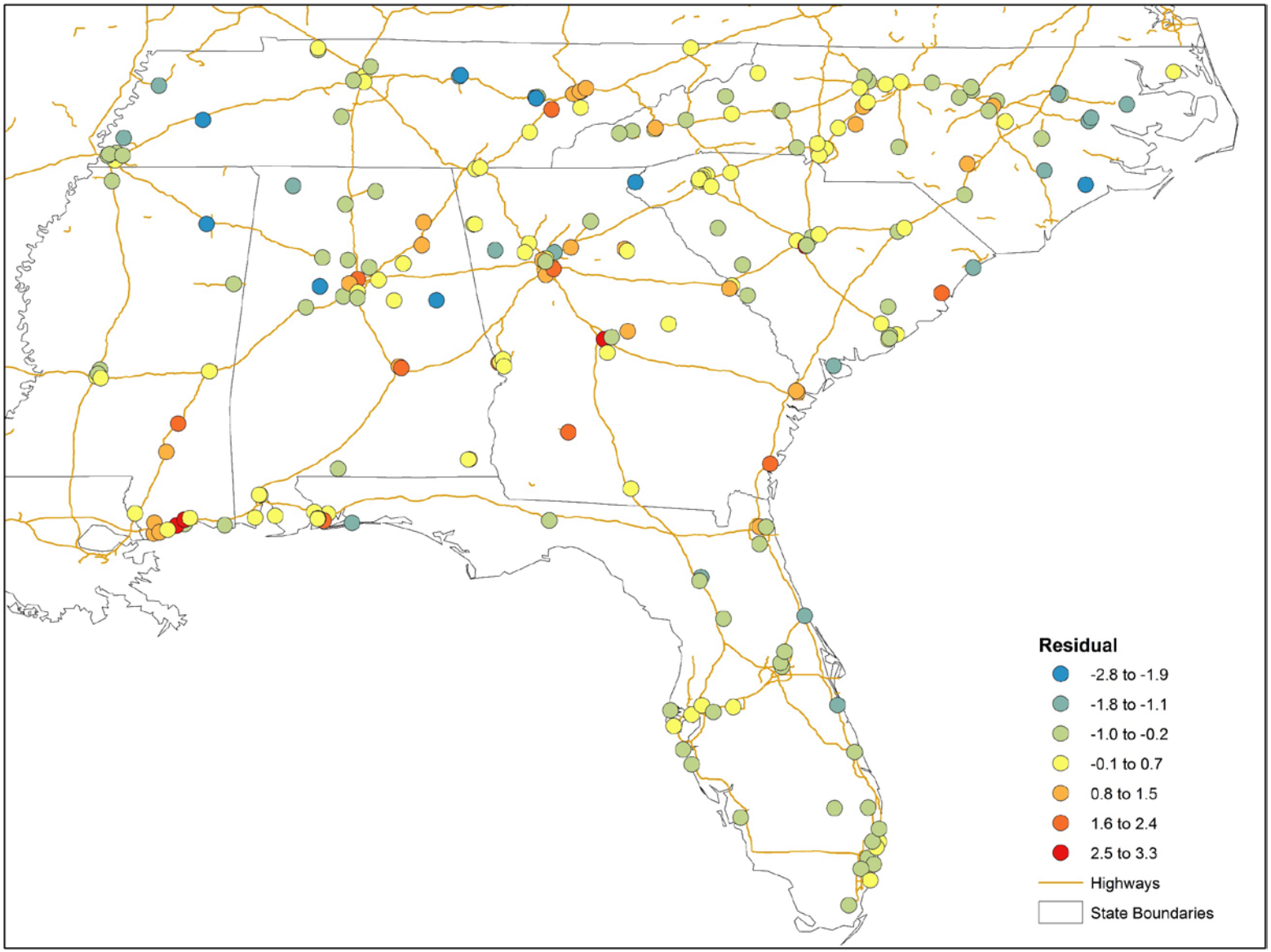

Figure 4. Residual Map